Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset

Ricardo Malheiro, Renato Panda, Paulo Gomes, Rui Paiva

CISUC – Centre for Informatics and Systems of the University of Coimbra
{rsmal, panda, pgomes, ruipedro}@dei.uc.pt

Abstract. This research addresses the role of audio and lyrics in the music emotion recognition. Each dimension (e.g., audio) was separately studied, as well as in a context of bimodal analysis. We perform classification by quadrant categories (4 classes). Our approach is based on several audio and lyrics state-of-the-art features, as well as novel lyric features. To evaluate our approach we create a ground-truth dataset. The main conclusions show that unlike most of the similar works, lyrics performed better than audio. This suggests the importance of the new proposed lyric features and that bimodal analysis is always better than each dimension.

Keywords: Bimodal Analysis, Music Emotion Recognition

1 Introduction

Music emotion recognition (MER) is gaining significant attention in the Music Information Retrieval (MIR) scientific community. In fact, the search of music through emotions is one of the main criteria utilized by users [1]. Most of early-stage automatic MER systems were based on audio content analysis (e.g., [2]). Later on, researchers started combining audio and lyrics, leading to bi-modal MER systems with improved accuracy (e.g., [3]).

The relations between emotions and music have been a subject of active research in music psychology for many years. Different emotion paradigms (e.g., categorical or dimensional) and taxonomies (e.g., Hevner, Russell) have been defined and exploited in different computational MER systems. Russell's circumspect model [4], where emotions are positioned in a two-dimensional plane comprising two axes, designated as valence (V) and arousal (A), as illustrated in Figure 1, is one of the well-known dimensional models. According to Russell [4], valence and arousal are the "core processes" of affect, forming the raw material or primitive of emotional experience. We use in this research a categorical version of this Russell's model, so we consider that a sentence belongs to quadrant 1 if both dimensions are positive; quadrant 2 if V is smaller than 0 and A is bigger than 0; quadrant 3 if both dimensions are negative and quadrant 4 if V is bigger than 0 and A is smaller than 0. The main emotions associated to each quadrant are illustrated in Figure 1.

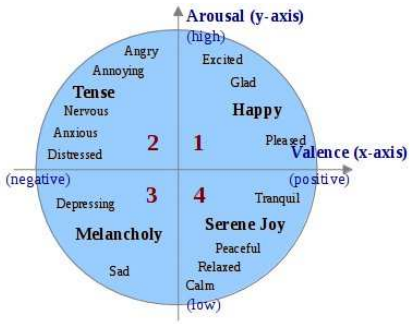

Fig 1. Russell's circumspect model (adapted from [4]).

Our goal is to find the best possible models for both dimensions (audio and lyrics) in a context of emotion recognition, using the Russell's model. To accomplish the goal we decided to construct a dataset manually annotated from the audio and from the lyrics. So, the annotators have been told explicitly to ignore the audio during the annotations of the lyrics to measure the impact of the lyrics in the emotions and do the opposite for the creation of the audio dataset. This approach is used by other researchers pursuing the same goals [5]. Then, we fused both dimensions and performed a bimodal analysis. For this study we use for both dimensions (audio and lyrics) almost all the state-of-the-art features that we are aware of, as well as new lyric features proposed by us [6].

2 Methods

2.1 Dataset Construction

To construct our ground truth, we started by collecting 200 song lyrics and the corresponding audio (30-sec audio clips). The criteria for selecting the songs were the following: Several musical genres and eras; Songs distributed uniformly by the 4 quadrants of the Russell emotion model.

The annotation of the dataset was performed by 39 people with different backgrounds. Each annotator classified, for the same song, either the audio or the lyric. During the process, people should: Identify the basic predominant emotion expressed by the audio / lyric (if the user thought that there was more than one emotion, he/she should pick the predominant one); Assign values (between -4 and 4 with a granularity of one unit) to valence and arousal.

For both, audio and lyrics dataset, the arousal and valence of each song were obtained by the average of the annotations of all the subjects. We obtained an average of 6 and 8 annotations respectively for audio and lyrics. To improve the consistency of the ground truth, the standard deviation (SD) of the annotations made by different subjects for the same song was evaluated. Using the same methodology as in [7], songs with an SD above 1.2 were excluded from the original set. As a result, the final audio dataset contains 162 audio clips (quadrant 1 (Q1) – 52 songs; quadrant 2 (Q2) – 45; quadrant 3 (Q3) – 31 and quadrant 4 (Q4) – 34), while the final lyrics dataset contains 180 lyrics

(Q1 – 44 songs; Q2 – 41; Q3 – 51 and Q4 – 44). Finally, the consistency of the ground truth was evaluated using Krippendorff's alpha [8], a measure of inter-coder agreement. This measure achieved, in the range -4 up to 4, 0.69 and 0.72 respectively for valence and arousal. This is considered a substantial agreement among the annotators. As for the lyrics the measure achieved 0.87 and 0.82 respectively for valence and arousal. This is considered a strong agreement among the annotators.

The size of the datasets is not too large, however we think that is acceptable for experiments and is similar to other datasets manually annotated (e.g., [7] has 195 songs).

Based on the lyrics and audio datasets, we created a bimodal dataset. We considered that a song (audio + lyrics) is a valid song to integrate this bimodal dataset, if the song belongs simultaneously to the audio and lyrics datasets and in both datasets the sample belongs to the same quadrant, i.e., we can only consider songs in which the classification (quadrant) for the audio sample is equal to the classification for the lyric sample.

So we started from a lyrics dataset containing 180 samples and an audio dataset containing 162 clips, obtaining a bimodal dataset containing 133 songs (with audio and lyrics): 37 songs for Q1 and Q2, 30 for Q3 and 29 for Q4.

2.2 Feature Extraction

In musical theory, the basic musical concepts and characteristics are commonly grouped under broader distinct elements such as rhythm, melody, timbre and others [7]. In this work, we organize the available audio features under these same elements. A total of 1701 features were extracted.

As for lyric features, we used state-of-the-art features such as: bag-of-words (BOW) – unigrams, bigrams and trigrams – associated or not to a set of transformations, e.g., stemming and stop-words removal; part-of-speech (POS) tagging[1] followed by a BOW analysis; 36 features representing the number of occurrences of 36 different grammatical classes in the lyrics (e.g., number of adjectives). We also used all the features based on existing frameworks like Synesketch (8 features), ConceptNet (8 features), LIWC (82 features) and General Inquirer (182 features). In addition to the previous frameworks, we use features based on known dictionaries such as DAL (Dictionary of Affect in Language) [9] and ANEW (Affective Norms for English Words) [10]. Finally, we propose new features: Slang presence, which counts the number of slang words from a dictionary of 17700 words; Structural analysis features, e.g., the number of repetitions of the title and chorus, the relative position of verses and chorus in the lyric; Semantic features, e.g., dictionaries personalized to the employed emotion categories.

3 Experimental Results

We conduct one experiment which is classification by quadrants (4 categories – Q1, Q2, Q3 and Q4). We use Support Vector Machines (SVM) [11] algorithm, since, based

[1] They consist in attributing a corresponding grammatical class to each word

on previous evaluations, this technique performed generally better than other methods. The classification results were validated with repeated stratified 10-fold cross validation (with 20 repetitions) and the average obtained performance (F-Measure) is reported.

We construct first, both for audio and lyrics, the best possible classifiers. We apply, for each one of the dimensions, feature selection and ranking using the ReliefF algorithm [12]. Next, we combine the best features of audio and lyrics and construct, using the same prior terms, the best bimodal classifier.

We can see in the following table (Table 1) the performance of the best model for lyrics, audio and for the combination of the best lyric and audio features. The fields *#Features, Selected Features and FM (%)* represent respectively the total number of features, the number of selected features and the F-measure score attained after feature selection. In the last line, the total number of bimodal features is the sum of selected lyrics and audio features.

Classification by Quadrants	#Features	Selected Features	FM (%)
Lyrics	1232	647	79.3
Audio	1701	418	72.6
Bimodal	1065	1057	88.4

Table 1. Best classification results by quadrants.

As can be seen, the best lyrics-based model achieved better performance than the best audio-based model (79.3% vs 72.6%). This is not the more frequent pattern in the state of the art, where usually the best results are achieved with the audio. This happens for example in [3]. [13] is the only research, as far as we know, where lyrics performance supplants audio performance, but only for some few emotions. This suggests that our new lyric features have an important role for these results.

As we can see, both dimensions are important, since bimodal analysis improves significantly (at $p<0.05$ Wilcoxon Test) the results of the lyrics classifier (from 79.3% to 88.4%). Furthermore, the best bimodal classifier, after feature selection, contains almost all the features from the best classifiers of lyrics and audio (1057 features in 1065 possible features). This suggests the importance of the features from both dimensions.

4 Conclusions

This paper investigates the role of audio and lyrics separately as well as combined in a context of bimodal analysis in the MER process. We proposed a new ground truth dataset containing 200 songs (audio and lyrics) manually annotated according to Russell's emotion model. We considered for bimodal analysis, songs with audio and lyrics annotated in the same quadrant (133 songs). We performed classification by quadrants (4 categories). We used most of the audio and lyrics state of the art features, as well as novel lyrics features.

The main conclusions show that unlike most of the similar works in the state-of-the-art, lyrics performed better than audio. This suggests the importance of the new proposed lyric features. Another conclusion is that bimodal analysis is always better than each one of the dimensions separated.

Acknowledgment

This work was supported by CISUC (Center for Informatics and Systems of the University of Coimbra).

5 References

1. Vignoli, F. Digital Music Interaction concepts: a user study. In: 5th Int. Conf. on Music Information Retrieval, (2004)
2. Lu, C., Hong, J., Cruz-Lara, S. Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques. In: 3rd Taiwanese-French Conf. on Information Technology, (2006)
3. Laurier, C., Grivolla, J., Herrera, P. Multimodal music mood classification using audio and lyrics. In Proc. of the Int. Conf. on Machine Learning and App., (2008)
4. Russell, J. Core affect and the psychological construction of emotion. Psychol. Review, 110, 1, 145–172, (2003)
5. Li, J., Gao, S., Han, N., Fang, Z., Liao, J. Music Mood Classification via Deep Belief Network. In: IEEE International Conference on Data Mining Workshop, 1241-1245, (2015)
6. Malheiro, R., Panda, R., Gomes, P., Paiva, R. Classification and Regression of Music Lyrics: Emotionally-Significant Features. In: 8th International Conference on Knowledge Discovery and Information Retrieval, Porto, (2016)
7. Yang, Y., Lin, Y., Su, Y., Chen, H. A regression approach to music emotion recognition. In: IEEE Transactions on audio, speech, and language processing, 16, 2, 448–457, (2008)
8. Krippendorff, K. Content Analysis: An Introduction to its Methodology. In: 2nd edition, chapter 11. Sage, Thousand Oaks, CA, (2004)
9. Whissell, C. Dictionary of Affect in Language. In: Plutchik and Kellerman (Eds.) Emotion: Theory, Research and Experience, 4, 113–131, Academic Press, (1989)
10. Bradley, M., Lang, P. Affective Norms for English Words: Stimuli, Instruction Manual and Affective Ratings. Technical report C-1, The Center for Research in Psychophysiology, University of Florida, (1999)
11. Boser, B., Guyon, I., Vapnik, V. A training algorithm for optimal margin classifiers. In: 5th Ann. Workshop on Computational Learning Theory, 144–152, (1992)
12. Robnik-Šikonja, M., Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RreliefF. Machine Learning, 53, 1–2, 23–69, (2003)
13. Hu, X., Downie, J., Ehmann, A. Lyric text mining in music mood classification. In: 10th Int. Society for Music Information Retrieval Conf., Japan, 11–416, (2009)